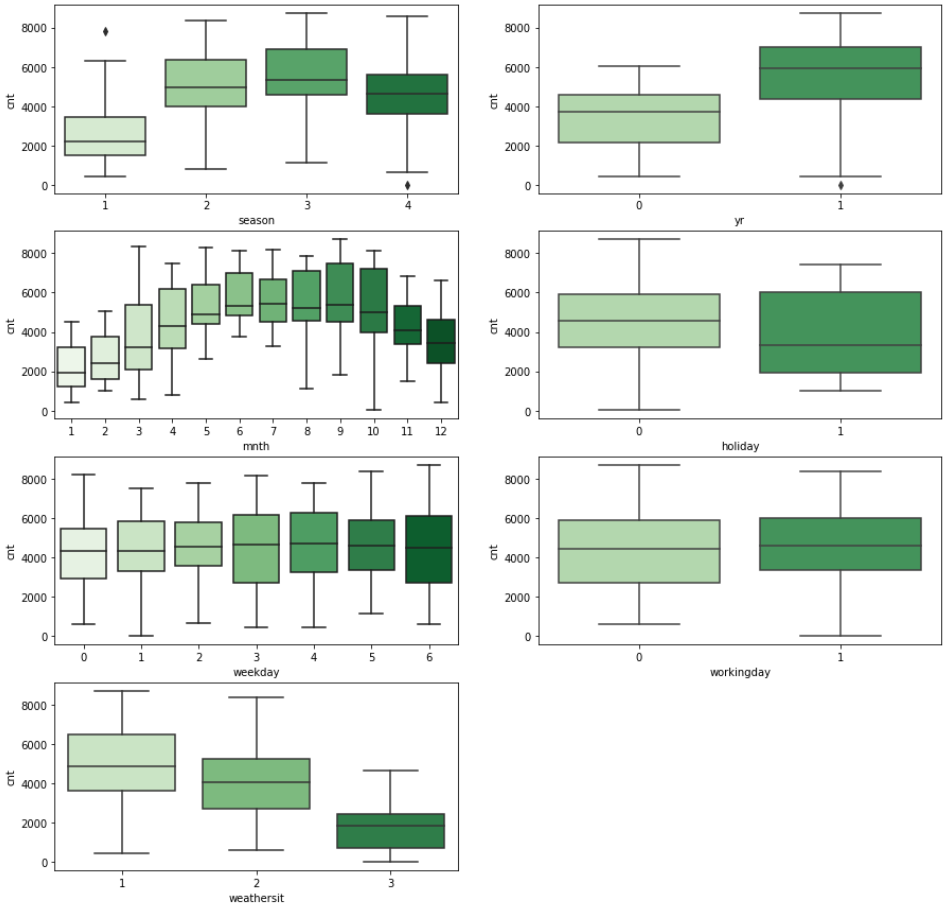
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Based on the Box Plot between all the categorical variables and dependent variable, we can infer the below points:

- Demand has been increasing year by year as business grows old
- Demand is continuously going up each month till September month and after that demand is decreasing
- Season fall has high demand for rental bikes
- On holidays, demand is low.
- Weekday has no major impact on demand
- Clear weather has highest demand compared to others



2. Why is it important to use drop_first=True during dummy variable creation?

drop_first=True, is important as it removes one column from the data frame while creating dummy variables, as removing that will actually not affect the overall analysis and model building as that column's importance can be explained by other remaining columns.

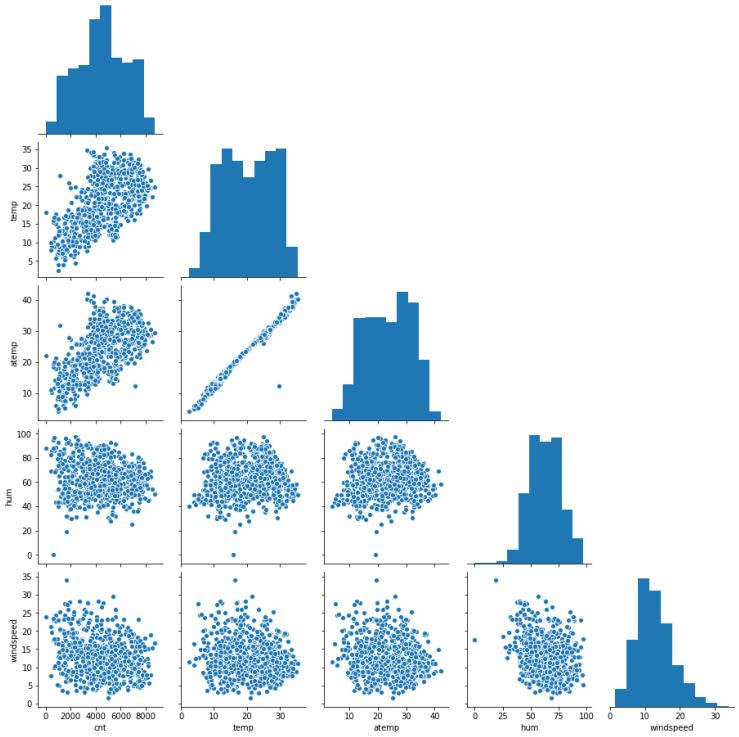
Consider, you have a column ‘gender’ that contains 3 variables- "Male", "Female", "Other". So a person is either "Male", or "Female", or "Other". Applying **drop_first=True** while creating dummy variables will actually create only two columns gender_male and gender_female only, so If they are not either of these 2, then their gender is "other".

Hence it reduces the correlations created among dummy variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Based on the Pair Plot between all the numerical variables and dependent variable, we can infer the below points:

- ‘atemp’ and ‘temp’ has the highest corelation with the target variable 'cnt'
- ‘temp’ and ‘atemp’ are hugely co-related with each other

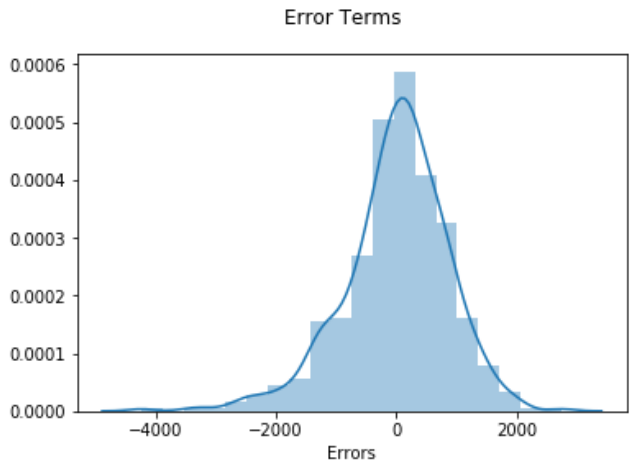


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Checked the below assumptions:

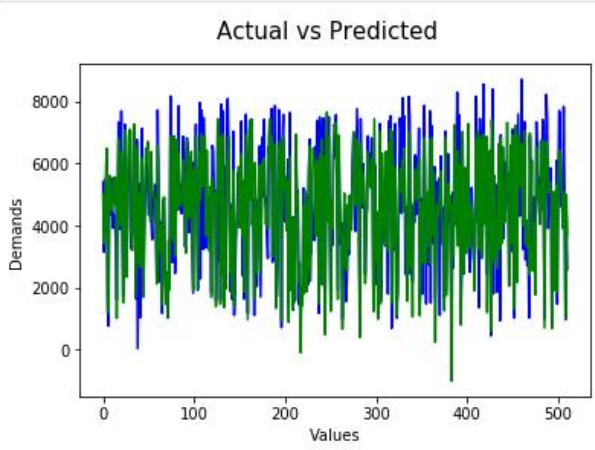
<u>Residual Analysis</u>

From below plot we can say that Error terms are normally distributed, here with mean ‘0’



Actual Vs Predicted Values (Linear relationship check**)**

Actual and Predicted Values are almost the same and are representing the same pattern



Multicollinearity Check

From the above VIF calculation we could see that there is no multicollinearity between the predictor variables, as all the values are within the range of below 5

	Features	VIF
3	windspeed	4.56
2	temp	4.10
0	yr	2.02
4	season_spring	1.67
5	season_winter	1.44
6	mnth_dec	1.22
7	mnth_sep	1.16
8	weathersit_Light_snowrain	1.09
1	holiday	1.03

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

Top 3 features contributing to towards the demand are below:

- September Month
- Temp
- Windspeed

General Subjective Questions

1. **Explain the linear regression algorithm in detail**

Linear regression is a model that estimates the relationship between one or more independent variable and one dependent variable using a straight line or plane.

If we use advertising spends as the predictor or independent variable, linear regression estimates the Sales.

It can be represented by y = B0 + B1*x1+B2*X2….

Where B0 is the bias coefficient and B1 and B2 are the coefficient (Simply the weightage) of X1 and X2

Ordinary Least Squares or Gradient Descent process is commonly used to calculate the coefficients of independent variables.

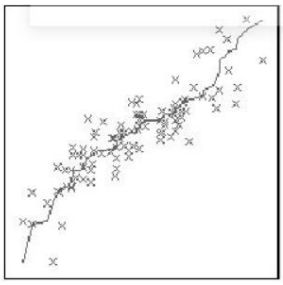
It can explain the kind of relationship between the predictor and independent variables:

House Price($)	YearsOld	Space(sq ft)	Owner
12000	5	1200	Sharath
2000	20	600	Raju
30000	2	4000	Unknown

Positive relationship:

When the regression line between the two variables moves in the same direction with an upward slope, the variables are said to be in a positive relationship.

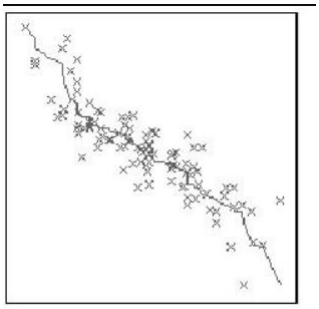
From the above example we can conclude that the house prices go up as the Space goes up.



Positive correlation

Negative relationship:
When the regression line between the two variables moves in the same direction with a downward slope, the variables are said to be in a negative relationship.

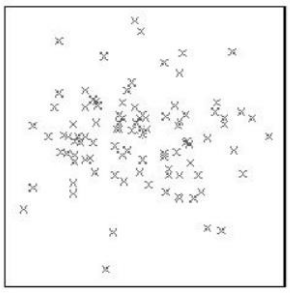
From the above example we can conclude that the house prices go down as the building becomes older.



Negative correlation

No relationship:
If the best fit line is flat (not sloped), it's assumed that there is no relationship among the variables

From the above example we can conclude that the house prices have no impact on the owner.



Zero correlation

2. Explain the Anscombe's quartet in detail.

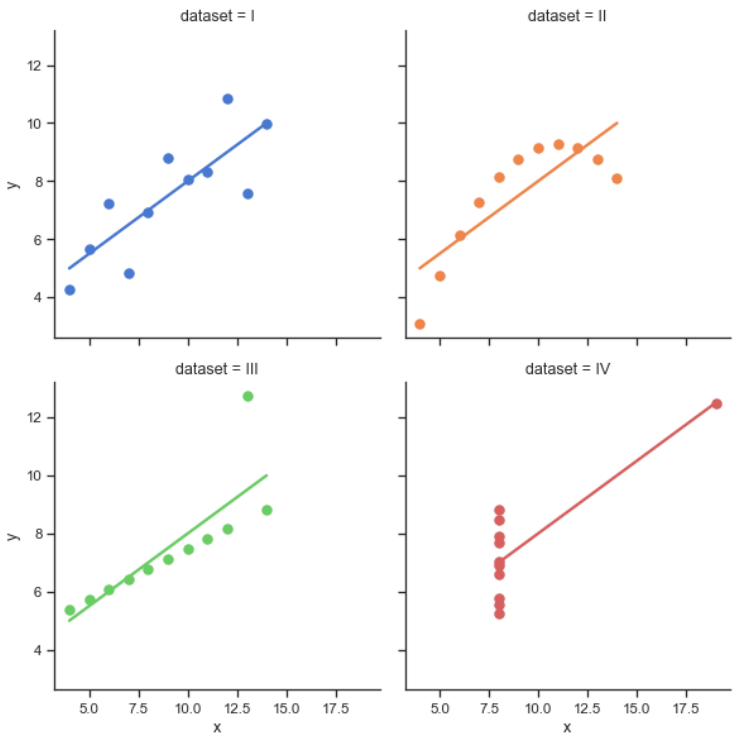
Anscombe's Quartet comprises four datasets, each containing x and y pairs. The important thing to note here about these datasets is that they have the same descriptive statistics. But story starts changing as soon as we graph or plot them.

Descriptive Statistics:

Set 1		Set 2		Set 3		Set 4	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
4	4.26	4	3.1	4	5.39	8	6.58
5	5.68	5	4.74	5	5.73	8	5.76
6	7.24	6	6.13	6	6.08	8	7.71
7	4.82	7	7.26	7	6.42	8	8.84
8	6.95	8	8.14	8	6.77	8	8.47
9	8.81	9	8.77	9	7.11	8	7.04
10	8.04	10	9.14	10	7.46	8	5.25
11	8.33	11	9.26	11	7.81	8	5.56
12	10.84	12	9.13	12	8.15	8	7.91
13	7.58	13	8.74	13	12.74	8	6.89
14	9.96	14	8.1	14	8.84	19	12.5

Though the summary statistics show that the means and the variances were identical for x and y across the groups, when plotted each graph tells a different story irrespective of their similar summary statistics.

Scatter Plots:



- **Dataset I**, appear to have clean and well-fitting linear models.
- **Dataset II**, represents a parabola shape and is not distributed normally.
- **Dataset III**, the distribution is linear, but it is heavily disturbed by an outlier.
- **Dataset IV**, shows that one outlier is enough to produce a high correlation coefficient.

Anscombe's Quartet shows that multiple data sets with many similar statistical properties can still be vastly different from one another when graphed. Additionally, it warns of the dangers of outliers in data sets while model building.

3. What is Pearson's R?

The Pearson's R is also known as Pearson correlation coefficient (r) measures the strength of the linear relationship between two variables.

It has a value between -1 to 1,

- **-1**, being a total negative linear correlation
- **0,** being no correlation,
- **+1,** being a total positive correlation.

Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

Example:

Using the below data, when the Pearson's R value is calculated, it's 0.952 which indicates a high correlation between both 'Experience' and 'Annual Income'

Experience	Annual Income ($)
1	10000
2	15000
3	25000
4	15000
5	40000
6	45000
7	52000
8	60000
Pearson's R	**0.952215823**

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

What is Scaling:
Feature scaling or simply scaling is a method used to standardize the range of independent variables or features of data to be used. In data processing, it is also known as data normalization or standardization. It is generally performed during the data pre-processing stage, before training models using machine learning algorithms.

The goal is to transform the data so that each feature is in the same range (e.g. between 0 and 1). This ensures that no single feature dominates the others, and makes training and tuning quicker and more effective

Few ways to do feature scaling:

- Normalization
- Standardization

Why is scaling performed:
For example, considering two features, weight and salary. weight is usually between 1 and 150 Kgs, while salary varies between 0 and 1 crore.
If we apply a machine learning algorithm to this dataset without feature scaling, the algorithm will give more weight to the salary feature since it has a much larger range. However, by rescaling both features to the range 0-1, we can give both features equal weight and improve the performance of our machine learning algorithm.

Feature scaling is performed when the dataset contains features that are highly
varying in magnitudes, units, and ranges.

Difference Between Normalized Scaling and Standardized Scaling:

Normalization is used when the data doesn't have Gaussian distribution whereas Standardization is used on data having Gaussian distribution.

Normalization:
It is the simplest method and consists of rescaling the range of features to scale the range in [0, 1]. The general formula for normalization is given as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Here, max(x) and min(x) are the maximum and the minimum values of the feature respectively.
Normalization is highly affected by outliers.

Standardization:
It is not bounded by range, feature standardization makes the values of each feature in the data have zero mean and unit variance. The general method of calculation is to determine the distribution mean and standard deviation for each feature and calculate the new data point by the following formula:

$$x' = \frac{x - \bar{x}}{\sigma}$$

Here, σ is the standard deviation of the feature, and $\bar{x}$ is the average of the feature.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

A variance inflation factor (VIF) provides a measure of multicollinearity among the independent variables in a multiple regression model.

VIF can be considered as a tool to identify the degree of multicollinearity between all the independent variables.

The dependent variable is the outcome that is being predicted by using the independent variables—which are the inputs into the model. Multicollinearity exists when there is a linear relationship, or correlation, between one or more of the independent variables or inputs.

Following should be taken into considerations:

- VIF equal to 1 = variables are not correlated
- VIF between 1 and 5 = variables are moderately correlated
- VIF greater than 5 = variables are highly correlated

The higher the VIF, the higher the possibility that multicollinearity exists, and further research is required. When VIF is higher than 10, there is significant multicollinearity that needs to be corrected.

VIF = infinity, tells that there is a huge correlation, it means that the variable is exactly linear combination of another independent variable being considered. If the independent variable can be explained perfectly by other independent variables, then it will have high or perfect correlation and its R-squared value will be equal to 1.

So, as per the calculations VIF = 1/(1-R Squared), which gives VIF = 1/0 which results in “infinity”. In this case the independent variable has to be dropped considering the domain knowledge.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Quantile - Quantile (Q-Q) plot, is a graphical tool that help us assess if a set of data comes from same distribution such as a Normal, Exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as distance, scale, and skewness are similar or different in the two distributions.

The power of Q-Q plots lies in their ability to summarize any distribution visually.

Importance of Q-Q plot:

- To fit a linear regression model, check if the points lie approximately on the line, and if they don’t, your residuals aren’t Gaussian and thus your errors aren’t either.
- When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If two samples do differ, it is also useful to gain some understanding of the differences.
- The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square.
- Many distributional aspects can be simultaneously tested.